# Case Studies and Regression Analysis

#### Jason Seawright

j-seawright@northwestern.edu

August 11, 2010

J. Seawright (PolSci)

Essex 2010

문에서 문어 .....

August 11, 2010

3

# Examples of Strengths and Weaknesses

J. Seawright (PolSci)

Essex 2010

August 11, 2010

▲ 글 ▶ ▲ 글 ▶ ... 글

DQC

## **Case Selection**

J. Seawright (PolSci)

Essex 2010



æ

590



Study the entire population.



Essex 2010



< ≣ ▶

E.

DQC

### Study the entire population.

2 Take a random sample.

J. Seawright (PolSci)

Essex 2010

<2.5 ≥ 2.2

August 11, 2010

#### Study the entire population.

- **2** Take a random sample.
- Follow some rule for deliberate case selection.

## Mill's Methods

J. Seawright (PolSci)

Essex 2010

August 11, 2010

イロト イヨト イヨト イヨト

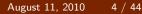


æ

#### • Method of Agreement



Essex 2010

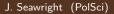


ヨトメヨト

Э

DQC

- Method of Agreement
- Method of Difference



Essex 2010



< ∃ >

æ

- Method of Agreement
- Method of Difference
- Etc..

< ∃ >



臣

## **Crucial Cases**

J. Seawright (PolSci)

Essex 2010

August 11, 2010

æ

590

5 / 44

イロト イヨト イヨト イヨト

 Theory assigns high likelihood to an outcome in a particular case that, for (all, or most, or a major) competing theory has low likelihood.

## **Regression Analysis**

- In a survey of 1000 articles in 10 leading political science journals, 49% used statistics (Bennett, Barth, and Rutherford 2003).
  - Presumably, most of them involve some variant of regression.

## **Regression Analysis**

- In a survey of 1000 articles in 10 leading political science journals, 49% used statistics (Bennett, Barth, and Rutherford 2003).
  - Presumably, most of them involve some variant of regression.
- A search in JStor for the word "regression" finds 10,404 relevant articles.

A 3 3 4

## **Regression Analysis**

 Almost no matter what you work on, you will have to interact with regression-based studies.

J. Seawright (PolSci)

Essex 2010

<2.5 ≥ 2.2

7 / 44

August 11, 2010



#### • Case-selection rules:



Essex 2010



Image: Image:

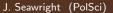
æ

DQC



#### • Case-selection rules:

• Random sampling



Essex 2010



물에 귀절에 다

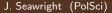
臣

DQC



#### • Case-selection rules:

- Random sampling
- Typical cases



Essex 2010

물에 귀절에 다

August 11, 2010

臣

DQC

#### • Case-selection rules:

- Random sampling
- Typical cases
- Diverse cases

< ≣ ▶

臣



#### • Case-selection rules:

- Random sampling
- Typical cases
- Diverse cases
- Extreme cases

< ∃ >



臣

#### • Case-selection rules:

- Random sampling
- Typical cases
- Diverse cases
- Extreme cases
- Deviant cases

J. Seawright (PolSci)

Essex 2010

August 11, 2010

→ ∃ →

围

#### • Case-selection rules:

- Random sampling
- Typical cases
- Diverse cases
- Extreme cases
- Deviant cases
- Influential cases

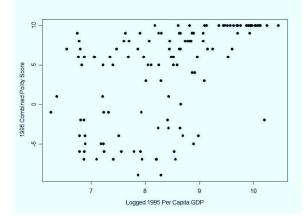
Essex 2010

August 11, 2010

#### • Case-selection rules:

- Random sampling
- Typical cases
- Diverse cases
- Extreme cases
- Deviant cases
- Influential cases
- Most-similar cases

## Running Example



J. Seawright (PolSci)

Essex 2010

August 11, 2010

臣

DQC



#### $\mathsf{Typicality}_i = -\mathsf{abs}[y_i - \mathsf{E}(y_i | x_{1,i}, x_{2,i}, \dots, x_{k,i})] \quad (1)$

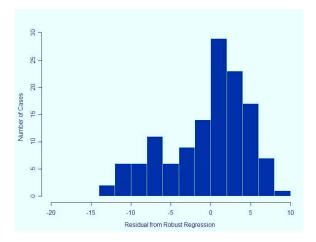
J. Seawright (PolSci)

Essex 2010

August 11, 2010

◆□ > ◆□ > ◆臣 > ◆臣 > □臣 = のへで

# Typical Cases



J. Seawright (PolSci)

Essex 2010

Э August 11, 2010

P

< ∃ >

æ

DQC



$$\mathsf{Extremity}_i = |\frac{x_i - \bar{x}}{s}|$$

J. Seawright (PolSci)

Essex 2010

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

(2)

æ

590



#### $Deviantness_i = -Typicality_i$

J. Seawright (PolSci)

Essex 2010

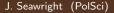
August 11, 2010

▶ ★ 문 ▶ ★ 문 ▶

(3)

= 990

 Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.



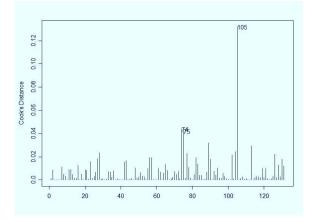
Essex 2010

August 11, 2010

- Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.
- A Cook's distance score of 1 or more usually is regarded as representing substantial influence.

August 11, 2010

## Influential Cases



J. Seawright (PolSci)

Essex 2010

August 11, 2010

æ

æ

200

 Matching techniques are an automated way of finding most similar cases.

Essex 2010

▶ ★ ≣ ▶

16 / 44

August 11, 2010

## Measurement Error in Y

 $Y_i^* = Y_i + \delta_{Y,i}$ 

J. Seawright (PolSci)

Essex 2010

August 11, 2010

回 と く ヨ と く ヨ と

= 990

## Measurement Error in Y

#### $Y_i^* = Y_i + \delta_{Y,i}$ Random Sampling

J. Seawright (PolSci)

Essex 2010

문에 비원이다.

August 11, 2010

臣

## Measurement Error in Y

Typical/Deviant Cases:  
$$e_i = Y_i - \mathbb{H}_{i,\cdot}Y + \delta_{Y,i}$$

J. Seawright (PolSci)

Essex 2010

August 11, 2010

æ

590

18 / 44

・ロト ・回ト ・ヨト ・ヨト

Influential Cases Strategy: Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

August 11, 2010

#### Extreme Cases: $Y_i^* = Y_i + \delta_{Y,i}$

J. Seawright (PolSci)

Essex 2010

< 2 → < 2 → ...

August 11, 2010

æ

DQC

#### Most-Similar Cases

J. Seawright (PolSci)

Essex 2010

문에서 문어?

August 11, 2010

DQC

21 / 44

臣

#### Most-Similar Cases Most-Different Cases

J. Seawright (PolSci)

Essex 2010

< ≣ ▶

August 11, 2010

臣

 $X_i^* = X_i + \delta_{X,i}$ 



Essex 2010



→ 御 → → 注 → → 注 → 三 注

590

#### $X_i^* = X_i + \delta_{X,i}$ Random Sampling

J. Seawright (PolSci)

Essex 2010



★ E ► < E ► </p>

臣

DQC

Typical/Deviant Cases:  
$$e_i = Y_i - X_i \hat{eta}^* - \delta_{X,i} \hat{eta}^*$$

J. Seawright (PolSci)

Essex 2010

・ロト ・回ト ・ヨト ・ヨト

August 11, 2010

E

590

#### Influential Cases Strategy: Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

August 11, 2010

#### Extreme Cases: $X_i^* = X_i + \delta_{X,i}$

J. Seawright (PolSci)

Essex 2010

August 11, 2010

æ

DQC

#### Most-Similar Cases

J. Seawright (PolSci)

Essex 2010

August 11, 2010

문에서 문어?

DQC

26 / 44

臣

#### Most-Similar Cases Most-Different Cases

J. Seawright (PolSci)

Essex 2010

< ∃ >

August 11, 2010

臣

#### $e_i = d_i + \gamma \tilde{Z}_i$ , where $\tilde{Z}_i = Z_I - \mathsf{E}(Z_i | X_i)$

J. Seawright (PolSci)

Essex 2010

August 11, 2010

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 三臣 - のへで

#### $e_i = d_i + \gamma \tilde{Z}_i$ , where $\tilde{Z}_i = Z_I - E(Z_i|X_i)$ Random Sampling

J. Seawright (PolSci)

Essex 2010

August 11, 2010

DQC

#### Typical/Deviant Cases: $e_i = d_i + \gamma \tilde{Z}_i$

J. Seawright (PolSci)

Essex 2010

문에 비원에 다

August 11, 2010

臣

DQC

#### Influential Cases Strategy: Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

August 11, 2010

### **Omitted Variables**

Extreme Cases:



Essex 2010



문에 비원에 다

E

DQC

Extreme Cases: For confounders, extreme on X may be a good strategy.

J. Seawright (PolSci)

Essex 2010

August 11, 2010

▶ ★ ≣ ▶

臣

- Extreme Cases:
- For confounders, extreme on X may be a good strategy. Extreme on Y maximizes:
- $\hat{Y}_i + d_i + \gamma \tilde{Z}_i$

- ▲ 臣 ▶ - -

30 / 44

August 11, 2010

### **Omitted Variables**

#### Most-Similar Cases

J. Seawright (PolSci)

Essex 2010

문에서 문어?

August 11, 2010

E

DQC

#### Most-Similar Cases Most-Different Cases

J. Seawright (PolSci)

Essex 2010

▶ < ≣ ▶

August 11, 2010

臣

DQC

## Pathway Variables

$$W_i = \nu + \mu X_i + \omega_i$$
  
$$Y_i = \alpha + \tau W_i + \sigma_i$$

J. Seawright (PolSci)

Essex 2010



æ

590

・ロト ・回ト ・ヨト ・ヨト

 $W_i = \nu + \mu X_i + \omega_i$   $Y_i = \alpha + \tau W_i + \sigma_i$ Random Sampling

J. Seawright (PolSci)

Essex 2010

★ E ► < E ► </p>

August 11, 2010

臣

# Pathway Variables

#### Typical/Deviant Cases: $e_i = \tau \omega_i + \sigma_i$

J. Seawright (PolSci)

Essex 2010

★ E ► < E ►</p>

August 11, 2010

æ

DQC

#### Influential Cases Strategy: Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

August 11, 2010

# Pathway Variables

Extreme Cases:



Essex 2010



★ E ► < E ►</p>

æ

590

Extreme Cases:  $W_i = \nu + \mu X_i + \omega_i$ 

J. Seawright (PolSci)

Essex 2010

▶ ★ E ▶ ★ E ▶

August 11, 2010

= 990

Extreme Cases:  $W_i = \nu + \mu X_i + \omega_i$ Extreme on Y maximizes:  $Y_i = \alpha + \tau W_i + \sigma_i$ 

J. Seawright (PolSci)

Essex 2010

August 11, 2010

문에 비원에 다

3

# Pathway Variables

#### Most-Similar Cases

J. Seawright (PolSci)

Essex 2010

August 11, 2010

< 2> < 2> < 2> <

æ

590

# Pathway Variables

#### Most-Similar Cases Most-Different Cases

J. Seawright (PolSci)

Essex 2010

August 11, 2010

물 🖌 🛪 물 🕨

臣

DQC

# Summary: Analytic Arguments

|            | Deviant | Influential | Ext. X | Ext. Y |
|------------|---------|-------------|--------|--------|
| Error in Y | Good    | Mixed       | Poor   | Good   |
| Error in X | Mixed   | Mixed       | Good   | Poor   |
| Confound   | Good    | Mixed       | Mixed  | Good   |
| Pathway    | Good    | Mixed       | Good   | Mixed  |

Essex 2010

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ 日

August 11, 2010

590

Simulate case selection for the same problem 10,000 times.



医外球 医外口

August 11, 2010

臣

DQC

Simulate case selection for the same problem 10,000 times.

• Analysis of presidential vote shares and the economy in Latin America, 1980-2000.

August 11, 2010

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.

August 11, 2010

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.
- 2 SD Rule

August 11, 2010

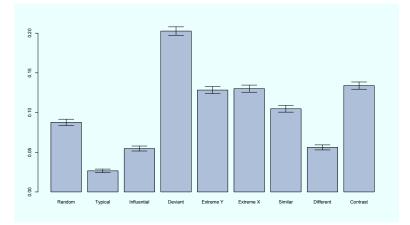


Figure: Case Selection for Finding Confounder.

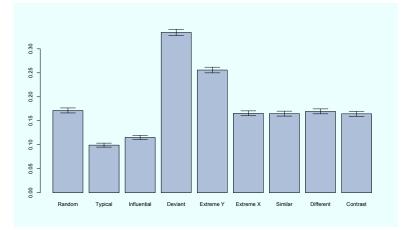
J. Seawright (PolSci)

Essex 2010

August 11, 2010

<= ∃ →

臣



#### Figure: Case Selection for Other Causes.

J. Seawright (PolSci)

Essex 2010

August 11, 2010

문에 비원에 다

臣

DQC

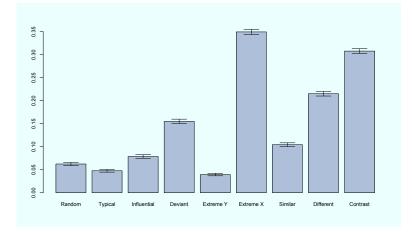


Figure: Case Selection for Exploring Mechanisms.

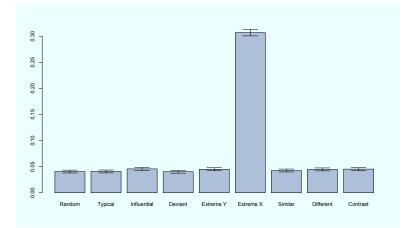
J. Seawright (PolSci)

Essex 2010

August 11, 2010

문 🕨 문

DQC



#### Figure: Case Selection for Error in X.

J. Seawright (PolSci)

Essex 2010

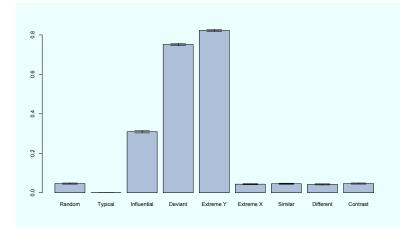
August 11, 2010

▶ ★ 臣 ▶ ....

200

42 / 44

臣



#### Figure: Case Selection for Error in Y.

J. Seawright (PolSci)

Essex 2010

August 11, 2010

물 에 문 에 문 이 나는 것

Ð,

590

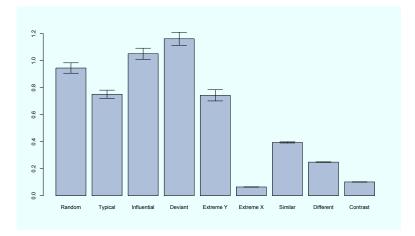


Figure: Case Selection for Estimating Overall Slope.

J. Seawright (PolSci)

Essex 2010

August 11, 2010

< ∃ >

DQC

#### Case-selection software in R

J. Seawright (PolSci)

Essex 2010

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

æ

590

Implement each case-selection technique for a data set of interest, or off my website. Be prepared to discuss what kinds of cases you get, and whether they seem on first glance to be useful, tomorrow.

J. Seawright (PolSci)

August 11, 2010